Quantifying the development of phraseological competence in L2 English writing: An automated approach

Yves Bestgen and Sylviane Granger

Centre for English Corpus Linguistics

Université catholique de Louvain

Full reference
Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing, 26, 28-41.

Link to publisher version: http://dx.doi.org/10.1016/j.jslw.2014.09.004

Authors' Note

Correspondence concerning this article should be addressed to Sylviane Granger, Centre for English Corpus Linguistics, Université catholique de Louvain, 1 Place Blaise Pascal, B-1348 Louvain-la-Neuve, Belgium.

Abstract

Based on the large body of research that shows phraseology to be pervasive in language, this study aims to assess the role played by phraseological competence in the development of L2 writing proficiency and text quality assessment. We propose to use CollGram, a technique that assigns to each pair of contiguous words (bigrams) in a learner text two association scores (mutual information and t-score) computed on the basis of a large reference corpus, the Corpus of Contemporary American English. Applied to the Michigan State University Corpus of second language writing, CollGram shows a longitudinal decrease in the use of collocations made up of high-frequency words that are less typical of native writers. It also shows that the mean MI scores of the bigrams used by L2 writers are positively correlated with the quality of the essays, while there is a negative correlation between the quality of the texts and the proportion of bigrams that were absent in the reference corpus, most of which were shown to be erroneous. The conclusion discusses the marked differences in the effects revealed by the longitudinal and pseudolongitudinal analyses, the limitations of the study and some potential implications for the teaching and assessment of second language writing.

Keywords: phraseology, n-gram, collocation, association measure, L2 learner corpus, writing assessment

Quantifying the development of phraseological competence in L2 English writing: An automated approach

Second language acquisition (SLA) has traditionally focused more on how L2 learners acquire morphology and grammar than lexis:

> the focus has been on how learners acquire grammatical sub-systems, such as negatives or interrogatives, or grammatical morphemes such as plural {s} or the definite or indefinite articles. Research has tended to ignore other levels of language. A little is known about L2 phonology, but almost nothing about the acquisition of lexis. (Ellis, 1985, p. 5)

Although the situation has started to change in recent years, lexical indices of language development are still less frequently used than syntactic measures such as T-unit length or percentage of error-free T-units. In other fields, however, lexis has come to occupy a central position. Corpus linguistics, for example, is largely lexical, probably because of the ease with which lexical items and lexico-grammatical patterns can be extracted, sorted and analysed. In the field of foreign language teaching, Lewis's (1993) 'Lexical Approach' which is based on the idea that "language consists of grammaticalized lexis, not lexicalized grammar", has led to a growing lexicalization of the teaching syllabus. The notion of lexis that underlies these approaches is phraseological; in other words, it goes beyond the study of single words to include a wide range of multi-word units. The field of phraseology, that is "the study of the structure, meaning and use of word combinations" (Cowie, 1994, 3168), has undergone a profound transformation in recent years. Long confined to the fringes of language study, it is now moving centre stage. There is growing recognition that besides being governed by grammatical and semantic rules, language production also largely relies on pre-patterned segments, a tendency that Sinclair (1991) has termed the 'idiom principle', in opposition to the 'open choice principle', and defined as follows: "the principle of idiom is

that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110). Corpus linguistic tools and methods have helped uncover a much wider range of word combinations than has previously been analysed: besides traditional units such as idioms (*to spill the beans*), compounds (*red tape*) or phrasal verbs (*give up*), which are characterized by a high degree of syntactic fixedness and semantic non-compositionality, corpus techniques have brought to light several types of sequences that stand out by their high degree of co-occurrence and recurrence rather than their fixedness or opacity. These include collocations, that is words that co-occur frequently within a short distance of each other in a text (Sinclair, 1991, p. 170), like *grow + old, turn + blue, dramatic + increase*, and lexical bundles, that is the most frequent recurring sequences of words in a register (Biber et al., 1999, Ch. 13), for example *you see what I mean* in conversation or *it should be noted* in academic writing.

If, as demonstrated by corpus linguistic studies, phraseology is pervasive in language, it is essential to study its role in L2 writing development. As pointed out by Li and Schmitt (2009), "learning to write well also entails learning to use formulaic sequences appropriately" and "L2 learners' failure to use native-like formulaic sequences is one factor in making their writing feel nonnative" (p. 86). More precisely, it has been shown that the L2 writers use less diverse formulaic sequences than native writers (De Cock, Granger, Leech, & McEnery, 1998) and overuse the ones they master best (Granger, 1998; Li & Schmitt, 2009).

Coxhead and Byrd (2007, pp. 134-135) advance three reasons that justify a stronger focus on formulaic sequences in L2 academic writing classes based on the analysis of corpus data: (1) using ready-made sequences is easier for students than composing sentences word by word; (2) formulaic sequences are defining markers of fluent academic writing; (3) being at the

boundary between lexis and grammar, formulaic sequences are much easier to detect on the basis of corpus data than through the analysis of individual texts.

The kinds of questions we need to address with respect to the role played by phraseology in L2 writing include the following: Do L2 writers use phraseological units? What types of units do they use? How does phraseological competence develop over time? To what types of difficulties do multiword units give rise? Are phraseological errors due to transfer from the learners' mother tongue? A wide range of studies have attempted to answer these questions in recent years (for an overview, see Paquot & Granger, 2012; Ebeling & Hasselgård, in press; Ellis, Simpson-Vlach, Römer, Brook O'Donnell, & Wulff, in press). A large number of these rely on computer learner corpora (i.e., large electronic collections of texts produced by foreign or second language learners), and make use of automatic techniques to extract multiword units. The *n*-gram method, which consists in extracting contiguous sequences of *n* words – two words for bigrams, three words for trigrams, etc. – is growing increasingly popular and has resulted in a large body of research on the use of lexical bundles by L2 writers. The data used is usually a combination of native and learner corpus data. Using a widely used method referred to as Contrastive Interlanguage Analysis (Granger, 1996), the learner corpus data is set against comparable native data with a view to uncovering the specificities of learner use, or against other samples of learner data in order to assess their degree of generalizability. A range of L2 English learner populations have been investigated in this way: French (De Cock et al., 1998), Lithuanian (Juknevičienė, 2009), Swedish (Groom, 2009), Japanese (Ishikawa, 2009) and Chinese (Chen & Baker, 2010), to cite just a few. Some studies compare written and spoken production (De Cock, 2000, 2007). Although the results of these studies are not directly comparable as they make use of different criteria to identify the relevant units, some general tendencies emerge: L2 writers rely on a more limited repertoire of lexical bundles than native writers; they overuse the bundles they

are familiar with, often calqued on similar sequences in their L1, and underuse many of the native-like bundles; they also prove to have difficulty with register, introducing speech-like bundles in their formal writing.

While these studies shed light on many aspects of the L2 phrasicon, the picture they present is largely static: "While valuable, this 'point-in-time' approach has difficulty illustrating the longitudinal development of formulaic language" (Li & Schmitt, 2009, p. 97). The reason for this is the lack of large longitudinal corpora of L2 writing, itself due to the time and effort needed to collect them. The issue of the development of phraseological competence is not altogether absent from studies of the L2 phrasicon, however, as some researchers have carried out pseudolongitudinal studies (i.e., a sub-category of cross-sectional studies that incorporates the proficiency dimension). As described by Gass and Selinker (2001), "[O]ne can use a cross-sectional design to create a pseudolongitudinal study. In such a study, the emphasis, like that of a longitudinal study, is on language change (i.e., acquisition), with data being collected at a single point in time, but with different proficiency levels represented" (pp. 32-33). By using this method, "[A] longitudinal picture can be then constructed by comparing the devices used by the different groups ranked according to their proficiency". This research design was used by Vidakovic and Barker (2010), who investigated the lexical development of L2 learners of English using written responses to Cambridge ESOL writing examinations across five proficiency levels. They extracted and analysed the highly frequent 4-word bundles used by the L2 writers at these five levels. Their analysis showed that lexical bundles were rarely used by the lowest proficiency writers. Learning conventionalized strings of words started at the elementary level but was found to be truly productive only at the upper intermediate and advanced levels, where bundles were the most numerous and diverse. The problem with their approach, however, is that it relies only on the frequency of the multiword units and pays no attention to the degree of

association within the units. As pointed out by several authors (e.g., Biber, Conrad, & Reppen, 1998; Evert, 2009; Hunston, 2002), the frequency of a sequence, though in itself an important feature, is not sufficient to identify authentic multiword units, as it does not take into account the frequency of the individual words in the sequence: "The fact that a sequence of words is above a certain frequency threshold does not necessarily imply either psycholinguistic salience or pedagogical relevance" (Simpson-Vlach & Ellis, 2010, p. 490). Very frequent words stand a much greater chance than rare words of occurring in numerous highly frequent sequences. This is convincingly demonstrated by Evert (2009, pp. 1224-1225) on the basis of the *is_to* bigram. With a frequency of 260 occurrences, this sequence is one of the most frequent bigrams in the 1-million word Brown corpus. However, this high frequency is not evidence of phraseological status, as both *is* and *to* are very frequent words. If words were randomly ordered in the corpus, thereby breaking any linguistic relation, we would expect 260 *is_to* bigrams, which is exactly the number observed.

In an effort to overcome this weakness, Durrant and Schmitt (2009) have designed a new approach that assigns to each sequence extracted from a corpus of L2 writing two well-established association measures computed on the basis of a large native reference corpus: Mutual Information (MI) (also called 'pointwise mutual information') and t-score (Church & Hanks, 1990; Evert, 2004; Hunston, 2002). Both measures compare how often a sequence of words appears in a corpus with how often it would be predicted to appear on the basis of the frequency of the words that compose it. Two association measures are required because each highlights a different type of collocation: MI, which tends to highlight word sequences made up of low-frequency words (such as *tectonic plates*), and t-score, which brings out those composed of high-frequency words (such as *long way*). Durrant and Schmitt's (2009) study focuses on one type of sequence (i.e., contiguous pairs of words made up of a modifier (adjective or noun) followed by a noun). The results show that, compared to native writers,

L2 writers of English tend to underuse collocations with high MI scores and overuse collocations with high t-scores. In a recent study, based on a sample of 223 essays extracted from the International Corpus of Learner English (Granger, Dagneaux, Meunier, & Paquot, 2009) and assessed for text quality by two professional raters, Granger and Bestgen (in press) demonstrated that the same differences can be observed between intermediate and advanced learners: intermediate learners tend to overuse high-frequency collocations and underuse lower-frequency, but strongly associated, collocations. The study uses the same method as Durrant and Schmitt to measure collocational strength but differs from it in two main ways: it uses an automated procedure to extract the sequences from a part-of-speech-tagged (POS-tagged) version of the learner corpus and it analyses the full range of bigrams rather than being restricted to modifier-plus-noun sequences. This latter aspect is particularly important, as modifier-plus-noun sequences are quite rare in a text (less than 7% of the total number of bigrams) and therefore provide a very limited picture of the L2 phrasal lexicon.

One key characteristic of Granger and Bestgen's study (in press) is that it is pseudolongitudinal, involving a comparison of L2 writers exhibiting different proficiency levels. While both longitudinal and pseudolongitudinal studies shed some light on the development of the L2 phrasicon, the two modes differ in one major respect: only longitudinal data traces the development of the same individual learners over a given period of time. It is therefore essential to apply phraseological indices to truly longitudinal data and this is exactly what we aim to do in this study. Thus, our main objective is to establish, using a longitudinal approach, whether phraseological competence in L2 writing, assessed here on the basis of the quantity and quality of bigrams, develops over time. Our second objective is to determine, by means of a cross-sectional approach, whether indices of phraseological competence correlate with the raters' judgements of essay quality. As the study is conducted

both longitudinally and pseudolongitudinally, this will help identify the respective contribution of each research design to the study of L2 writing development.

Data and Methodology

Overview of the Methodological Approach

The technique aims to assign to each bigram (i.e., any contiguous pair of words in the L2 texts) association scores computed on the basis of a reference corpus. In view of their collocational status, the resulting units are referred to as 'collgrams' to distinguish them from *n*-grams, whose collocational status is unspecified. The technique itself will be referred to as the CollGram technique or CollGram for short.

The next step consists in computing three measures to quantify the collocational strength of each text:

- the mean MI score, which measures collocations made up of infrequent words;
- the mean t-score, which measures collocations composed of very frequent words;
- the proportion of bigrams that are absent from the reference corpus and thus cannot be assigned any association score. These bigrams may be errors or creative combinations.

The first stage is to extract all the bigrams in each L2 text. Second, the bigrams are looked up in the reference corpus to determine their frequency and are assigned the two association scores (MI and t-score). In a third step, the three indices are computed for each L2 text. As is evident from this description, the CollGram method requires a reference corpus in order to assign association scores to each bigram. This corpus needs to have the following two characteristics: it must be large enough to allow for a precise estimation of the association scores and it must be as representative as possible of the target language (here, English) as it is used today. Two widely used corpora that meet these two criteria are the British National Corpus (BNC), a 100-million word collection of samples of written and

spoken language designed to represent a wide cross-section of British English from the latter part of the 20th century, and the Corpus of Contemporary American English (COCA), a balanced, 425-million word corpus of American English collected from 1990 to 2011.

CollGram relies on the following automated processes:

1. Tokenization: each learner text is tokenized and POS-tagged with a view to identifying proper names and punctuation marks. We used CLAWS7[1], which has a high degree of accuracy overall (96-97%) and has proved to perform better than other POS-taggers when handling learner data (Van Rooy & Schäfer, 2003).
2. Bigram extraction: all bigrams are extracted from each L2 text. Punctuation marks and any sequence of characters that does not correspond to a word interrupt the bigrams. The sequences *cars_now* and *now_he* in the excerpt "*illegally parked cars. Now, he came*" (see Table 2 for this example and several others) were therefore not extracted, but *illegally_parked*, *parked_cars* and *he_cam*e were. In addition, bigrams that contain a word identified by CLAWS as a proper name or a number are excluded from subsequent analyses.
3. Computation of association scores: each bigram is looked up in the reference corpus and assigned its MI and t-score computed by means of the formulas reported in Evert (2009, p. 1225). These two indices compare the observed frequency of a bigram in the reference corpus with the expected frequency computed on the basis of the frequency of the two component words. MI is a measure of strength of association, originating from information theory, which corresponds to the log-transformed ratio between the observed frequency of the bigram and its expected frequency. A highly positive value signals a collgram made up of words that are rarely found independently of each other. As a result, even an infrequent bigram can have a high

[1] http://ucrel.lancs.ac.uk/claws/

MI score if it comprises very rare words. The t-score, which derives from classical statistical testing and not from information theory, is a measure of certainty: it expresses the confidence we can have about the existence of an association between two words. It is computed by dividing the difference between the observed frequency of the collgram and the expected frequency by the square root of the observed frequency (see Church, Gale, Hanks, & Hindle (1991) for the derivation of this test). Compared to MI, it gives much more weight to the number of times a collgram has been observed. It therefore prioritizes frequently occurring collgrams that are, in essence, made up of frequent words. It is important to note that these two association scores take into account the order of words in the collgram. As a result, a highly collocational sequence like *for example* does not obtain the same scores as the reverse sequence that is found in the sentence *He set a good example for the rest of us*. Obviously, these two association indices can be computed only if the bigram occurs in the reference corpus. If it fails to occur, the bigram is included in the 'absent from corpus' category.

4. Computation of collgram profiles: three indices – mean MI, mean t and proportion of absent bigrams – are used to draw the collgram profiles of L2 texts and relate them to the development of L2 writing proficiency. They are computed on the basis of all the bigrams present in the learner texts (tokens) as well as all the different bigrams present in each text (types). In counts based on bigram types, even if a learner uses a bigram several times, the bigram is only counted once. This score therefore gives greater weight to the diversity of bigrams present in a text.

Learner Corpus

The learner corpus used for the study is the Michigan State University (MSU) corpus of English as a Second Language writing made up of 171 essays written by 57 university-age

learners of English (see the introduction to this issue for details). Each essay had been evaluated twice by the same two expert raters at several months' interval on the basis of two different analytic scales, which included the following criteria: content, organization, vocabulary, language and mechanics. Like the other authors in this special issue, we opted for three measures of essay quality from the revised analytic scale described in the introductory section by Connor-Linton and Polio: the combined score computed over the five scales and the Language Use and Vocabulary specific scales. These three scores were obtained by averaging the two raters' assessments. The combined score is the most reliable score available, achieving an inter-rater correlation coefficient (Pearson) of r=0.88. It encompasses a large variety of text quality dimensions: content, organization, vocabulary, language and mechanics. Because some of these dimensions are not directly related to the phraseological competence that CollGram is supposed to tackle, we also selected the Language Use and Vocabulary subscales, whose reliability, though weaker, is adequate, since the inter-rater correlation coefficient is r=0.77 for Language Use and 0.76 for Vocabulary. The descriptors of the Language Use subscale mainly cover morphological and grammatical aspects (word order, morphological errors). The Vocabulary subscale covers lexical aspects like vocabulary sophistication, lexical errors and idiomatic lexical use. Comparing the results for these two scales will shed some light on the importance of linguistic components other than lexical for the automated measurement of phraseological competence.

Several pretreatments were applied to the learner texts before the bigram extraction stage. Spelling errors were removed to be able to group instances of the same word pair that differed only in one or two minor spelling errors (e.g., *private correspondance* and *private correspondence* were counted as two occurrences of the bigram *private correspondence*). We normalized words only when there was no doubt about the targeted form. This involves changes such as erroneous doubling of consonants (*appartment*), omission of a letter

(*completly*) or addition of a letter (*ridicoulous*). No attempt was made to normalize words like *documentals,* which could in principle stand for *document, documentation* or *documentary*. Contracted forms were automatically grouped with their corresponding full forms (*I've > I have*; *he's > he is*). Ambiguous contracted forms were easily disambiguated thanks to their POS tags (e.g., enclitic *'s* is tagged VBZ (*is*), VHZ (*has*) or POS (genitive) by CLAWS7). All the bigrams containing at least one word identified by CLAWS7 as a proper name or a number were excluded. The details of the learner corpus are provided in Table 1.

Insert Table 1 about here

Reference Corpus

The reference corpus we used is the Corpus of Contemporary American English (COCA), a very large and balanced corpus of American English[2]. The version we used contains more than 425 million words of text (20 million words each year from 1990-2011) and is equally divided among speech, fiction, popular magazines, newspapers and academic texts. Earlier studies by Durrant and Schmitt (2009) and Granger and Bestgen (in press) used the BNC, but we opted for COCA because American English is the dominant variety of English for the second language writers represented in the MSU learner corpus. At the time of the study, COCA was not distributed but it was possible to obtain frequency lists of all words and bigrams (just over 375 million) in the corpus. As the corpus was tokenized and POS-tagged with CLAWS, we were able to apply the same pretreatments as those used for the learner corpus (e.g., changing contracted forms into full forms).

[2] http://corpus.byu.edu/coca/

Results

This section is structured as follows. After illustrating the workings of CollGram on the basis of two marked-up passages, we provide a qualitative presentation of two key categories of bigrams in the MSU corpus: those that received the most extreme association scores and those that were found to be absent from the reference corpus. There follows a presentation of the results obtained from the two stages of the quantitative analysis carried out using, respectively, a longitudinal and a pseudolongitudinal approach.

Marked-up Passages

The two short passages included in Table 2 illustrate how the method works. Both passages contain 20 bigrams. The first one (essay 122; mean rating 70) is the excerpt that achieved the highest mean MI score (4.55) and the second (essay 296; mean rating 53.5) is the excerpt that is closest to a mean MI of 0 (0.008).

Insert Table 2 about here

The following observations can be made based on the table:

- Negative MI values (for example, *everything are*) correspond to bigrams that co-occur in the reference corpus less frequently than chance would predict. These bigrams can be erroneous combinations (e.g., *everything are*) or creative combinations (e.g., *ignominious award*).
- *Investigate illegally* is not present in the reference corpus and is therefore not assigned any association score. However, it is counted in the category of bigrams that are absent in the reference corpus.

- Punctuation marks: the presence of a punctuation mark interrupts bigrams. No score is computed and the sequence is not taken into account in the computation of absent bigrams. Note that the absence of a punctuation mark between *peaceful* and *Nothing* generates an additional bigram (which, in this particular case, is not present in the reference corpus).

Highest- and Lowest-scoring Bigrams

Table 3 lists the 50 highest- and lowest-scoring bigrams in the MSU learner corpus classified in decreasing order of the absolute value of the MI and t-score. The lowest-scoring bigrams are all bigrams that occur in the reference corpus less frequently than chance would predict.

It is immediately obvious from the left-hand side of Table 3 that the top-scoring sequences identified by MI and t-score, respectively, are of a completely different order. Many of the sequences that obtain top t-scores are composed of very frequent grammatical words (pronouns, prepositions, auxiliaries, determiners) and high-frequency lexical verbs (*think, want, get, say*). Many are of the type Preposition + Determiner (*of the, in the, on the, to the, in a, for a*) or Pronoun + Verb (*it was, I think, he was, I do, this is*). The list also contains some basic close-knit units like *out of, more than, the same* and *a lot*. The top-scoring bigrams identified on the basis of MI, however, all contain much less frequent words. The majority are compound-like units made up of Noun + Noun (*rocket launchers, ozone layer, personality traits*) or Adjective + Noun (*alcoholic beverage, acid rain, monetary fund*). The only two bigrams with high MI scores that contain verbs (*committed suicide* and *sun shines*) are very different from those identified by t-score.

The right-hand side of Table 3 shows that many of the lowest-scoring bigrams identified by MI are also identified by t-score. The shared bigrams (27 out of 50) mainly contain grammatical words used in erroneous combinations (*his my, a out, their are, a each,*

they is). The t-exclusive bigrams are of a similar order (*the some, the all*). The MI-exclusive bigrams, however, tend to contain lexical words (verbs, nouns, adjectives) used in grammatically or lexically erroneous combinations and provide evidence of a range of difficulties: article use (*a experience, a biggest*), bound preposition (*include of, filled of*), concord (*all person*), verb morphology (*to entered*), etc.

Insert Table 3 about here

Analysis of the Absent Category

One category that deserves particular attention is that of bigrams in the learner corpus that are absent in the reference corpus. Theoretically, these bigrams can be of two types: creative combinations, which are more likely to be used by advanced learners, and erroneous combinations, which will tend to be produced in greater quantity by less advanced learners. The latter are likely to be very similar to the negative MI bigrams illustrated above. To establish the proportion of erroneous bigrams in this category, we extracted a random sample of 200 bigrams that are absent from the reference corpus and analysed them in context. Roughly one-third (70 out of 200) are grammatically possible combinations that happen not to be present in the reference corpus (e.g., *analyzing similarities, brother-in-law graduated, convenient systems, ejected students*). The rest of the bigrams give an extremely rich picture of the wide range of problems learners encounter when they combine words. Table 4 illustrates some of the most frequent categories of errors: verb morphology (lines 1 and 2), number (lines 3 and 4), article use (lines 5 and 6), verb complementation (lines 7 and 8), preposition use (lines 9 and 10), confusion between grammatical categories (lines 11 and 12) and word coinage (lines 13 and 14).

Insert Table 4 about here

Longitudinal Analysis

For the longitudinal study, the analyses were conducted on the basis of the first and last essays written; the middle essays were disregarded. This approach, also adopted by Crossley and McNamara (this issue) and Bulté and Housen (this issue), finds its justification in the time period covered (i.e., one semester) a time period which, even in immersion, is relatively short for a longitudinal study (Storch, 2009).

Insert Table 5 about here

The means and standard deviations of the three measures (i.e.,MI, t-score and proportion of bigrams absent from the reference corpus) computed for both types and tokens are shown in Table 5. Using repeated measures ANOVAs, we found a statistically significant decrease in the t-score measure from time 1 to time 3 (Tokens: $F(1, 56) = 4.71$, $p < .05$, $\eta^2 = 0.08$; Types: $F(1, 56) = 6.53$, $p < .05$, $\eta^2 = 0.12$), but none of the other measures showed a significant longitudinal evolution. The smallest p-value for any other F test was obtained for the MI computed on the tokens and was larger than .10.

Pseudolongitudinal Analysis

For the pseudolongitudinal analysis, we computed the correlations between the three indices and the mean ratings of the essays for the three scales (combined, language use and vocabulary). The results are presented in Table 6. They show that the mean MI score is significantly linked to the rated quality of the text, while the correlations for the mean t-score are weak and never reach statistical significance. A significant correlation is observed for the proportion of absent bigrams. The correlation is negative: the more absent bigrams there are,

the lower the rating of the text. The correlations are higher for the language scale than for the combined scale and for the vocabulary scale, which is clearly the least well predicted scale by collgram scores. The latter result suggests that the CollGram technique includes a grammatical component in addition to the lexical component that is central to the vocabulary scale. There are few differences between types and tokens. However, the fact that the correlations are stronger for the types in MI may suggest that once a bigram has been used by a learner, its repetition in the same text does not provide any additional information that can be used to predict quality.

Insert Table 6 about here

The existence of statistically significant correlations between the rated quality of the texts and both the MI score and the proportion of absent bigrams suggests that a combination of the two indices might enhance quality prediction. To test this hypothesis, the two indices were introduced as predictors in a multiple regression with the language scale as dependent variable. The results highlight a weak improvement in the correlation for the combination of the two measures. The multiple correlation (i.e., the square root of the regression R-square) equals 0.48, a gain of 0.05 over the correlation between MI types and rated text quality in Table 6.

Discussion

The objective of the analyses reported above was to establish whether it was possible to use the CollGram technique to track the development of phraseological competence in L2 writing. The results are encouraging, but the study highlights marked differences in the effects revealed by the longitudinal and pseudolongitudinal analyses.

The longitudinal study showed a decrease in the number of high-frequency collgrams identified by t-score, but no significant difference in the number of low-frequency collgrams identified by the MI measure. This result may find its explanation in usage-based theories of language acquisition, which hold that L2 learners acquire constructions from the abstraction of patterns of form-meaning correspondence in their usage experience (Ellis et al., in press). Several studies on both L1 and L2 acquisition point to an acquisition sequence characterized by a progressive deconstruction of multiword units, from "low-level binary chunks like bigrams" (Ellis, 2003, p. 64) to more complex units like collocations and idioms: "[T]he typical route of emergence of constructions is from formula, through low-scope pattern, to construction" (Ellis, 2002, p. 143). The frequency distributions of each part of the formula play a key role in this process: formulae that occur frequently in learners' naturalistic or classroom environment are likely to be acquired first and develop quicker into low-scope patterns. The decrease in t-score bigrams may well reflect this process: as the two parts of the bigram are progressively encountered in a range of different contexts, learners start to use each of the words in more diversified contexts, a process that leads to the production of bigrams with lower t-scores. The lack of development of MI scores can be explained by the low frequency of the bigrams in the learners' input coupled with the short period of time covered. As pointed out by Ellis et al. (in press), "Many of the forms required for idiomatic use are relatively low frequency, and the learner thus needs a large input sample just to encounter them". One semester of L2 writing instruction and immersion in the target speech community may not have given learners enough opportunity to encounter the low-frequency bigrams.

The results of the pseudolongitudinal study present a completely different picture. No statistically significant correlation was observed between the quality of the essays and the mean t-scores. However, the mean MI scores of the bigrams used by learners were found to

be positively correlated with the quality of the essays. The absence of correlation for the t-score may be due to the lack of salience of bigrams made up of high-frequency words. It is reasonable to assume that bigrams made up of low-frequency words (i.e.,high MI collgrams) might be more readily noticed by raters and positively influence their judgement. Li and Schmitt (2009, p. 96) seem to suggest this when they state that less frequent, strongly associated collocations identified by means of the MI are "the type of item which is likely to be highly salient for native speakers". The significant correlation for the Absent category may be an indication of the importance of errors, in particular grammatical ones, in raters' judgements. Weltig's (2004) study of the effects of language errors on ESL raters' scores showed error density to have a strong and significant effect on both language and content scores. This is clearly in line with Polio and Shea's study (this issue), which showed very strong negative correlations between the number of errors per word and the holistic ratings of text quality. The correlations we have obtained are weaker, but they result from a fully automatic procedure which is bound to be less reliable than manual annotation carried out by two experts. In addition, as shown by the quantitative analysis of the absent bigrams, a certain number are creative combinations and can therefore be expected to be more frequent in the better essays. At this stage, CollGram cannot distinguish the absent bigrams that are creative from those that are erroneous.

Conclusions

Our study aimed to assess the role played by phraseological competence, assessed on the basis of the quantity and quality of bigrams, in the development of L2 writing proficiency and text quality assessment. Unlike the other contributions to this issue, which rely mainly on 'text-internal' measures (i.e., calculated solely on the basis of the learner text) our study relies on 'text-external' measures (Meara & Bell, 2001; Skehan, 2009), that are calculated on the basis of an external resource, namely a large corpus of texts covering a broad spectrum of

native language use. This external approach defines formulaicity in L2 in terms of its degree of similarity to native norms: "Thus we can also operationalize the formulaicity of L2 language by how well it uses the formulaic sequences and grammatico-lexical techniques of the norms of its reference genre" (Ellis et al., in press).

Using the CollGram technique, the analysis of the MSU corpus shows that the mean MI scores of the bigrams used by learners are positively correlated with the quality of the essays, while there is a negative correlation between the quality of the texts and the proportion of bigrams that were absent in the reference corpus, most of which were shown to be erroneous. On the other hand, the longitudinal analysis showed a statistically significant evolution of the t-score, thereby confirming the predictions that can be derived from previous studies by Durrant and Schmitt (2009) and Granger and Bestgen (in press).

One of the most unexpected findings of our study is the discrepancy between the longitudinal and pseudolongitudinal analyses. As pointed out by Connor-Linton and Polio in their introduction to this special issue, this finding is also highlighted in the other four articles, albeit less strikingly in the case of Friginal and Weigle (this issue). Like the other authors, we have provided an interpretation of this discrepancy based on the nature of the linguistic indices used in the study, but the fact that studies based on different techniques led to the same conclusion should perhaps prompt researchers to look for a more general interpretation. First and foremost, however, it is essential to assess the generalizability of this observation by applying the different measures to other corpora, which, like the MSU corpus, allow for both longitudinal and cross-sectional analyses. Unfortunately, this objective is difficult to achieve, as there is a severe shortage of corpora of this type, especially ones that reach the critical size needed to draw reliable conclusions. Ellis (2003) highlights the "need for larger-sampled SLA corpora which will allow detailed analysis of acquisition sequences" (p. 74) and considers filling this gap to be "an important research priority" (p. 73). When

collecting these corpora, care should be taken to clearly distinguish between second and foreign language learners. As degree and type of exposure are important determinants of language acquisition, one can reasonably expect second language learners to perform better than foreign language learners in some areas of language. As regards phraseological competence, Ellis et al. (in press) review a number of studies that suggest "a potential difference in formulaic use between ESL learners who are exposed to lots of naturalistic spoken language, and EFL learners who are not. Learning the usages that are normal or unmarked from those that are unnatural or marked requires a huge amount of immersion in the speech community". For example, Alsakran's 2011 study of collocation use by EFL and ESL Arabic-speaking learners reveals that the participants' learning environment has a strong effect on the acquisition of L2 collocations, with ESL learners obtaining significantly higher scores than EFL learners. A similar effect was found by Yamashita and Jiang (2010), who also underline the role of the learners' mother tongue for collocational development and use and conclude "that both L1 congruency and L2 exposure affect the acquisition of L2 collocations" (p. 647).

One of the major implications of our study is that phraseology plays a role in the development of L2 writing and should therefore be incorporated into the battery of linguistic indices used to track this development. The advantage of phraseological indices like collgrams is that they combine lexis and grammar. Ruegg, Fritz and Holland (2011) recently underlined raters' difficulties in assigning separate scores for the lexical and grammatical dimensions when assessing the writing quality of a text. Based on the idea that assessment criteria should reflect the "inextricable interwovenness" (Ruegg et al., 2011, p. 75) of lexis and grammar, they suggest merging the lexis and grammar scales into one single lexicogrammar scale. Our study suggests that it might be valuable to adopt a similar strategy

for the quantitative analysis of text quality indices. The CollGram technique, which extracts both lexical and grammatical sequences, is a first step in this direction.

Another area that could benefit from CollGram is foreign/second language teaching. In addition to the specific contribution it can make to the general assessment of L2 texts, CollGram can point to the collocations used by learners that are typically used by native speakers and those that are more rarely used by them, if at all. This type of information, highlighted by the qualitative analysis of the MSU data, has potential for L2 writing instruction. For example, the CollGram metric could be incorporated into a software tool like the ones recommended to L2 writing teachers by Coxhead and Byrd (2007), that automatically highlights in L2 texts the most and least native-like collocations (i.e., those that obtain the most positive and the most negative collgram scores), thereby helping teachers in marking students' work (see Coxhead and Byrd, 2007). It would also be possible to draw up lists of problematic collocations for learners with a specific mother tongue background and design pedagogical materials addressing L1-specific needs. As rightly noted by Henriksen (2013), the teaching of lexis is still largely single-word based: "Many teachers tend to focus on individual words (e.g., in glosses and tasks) and often lack useful materials for raising learners' awareness of collocations" (p. 41). CollGram could contribute to the gradual 'phrasing up' of L2 instruction advocated by a number of scholars, in particular Lewis (1993) and Nattinger and DeCarrico (1992).

On a more theoretical level, our study can be seen as a preliminary attempt to answer Skehan's (2009) recent call to supplement the three major components of L2 performance – complexity, accuracy and fluency – by measures of lexical performance: "lexis represents a form of complexity that has to be assessed in second language speech performance if any sort of complete picture is to be achieved" (p. 514). CollGram seems to be a promising candidate in this respect, as it is highly versatile: it incorporates the three traditional dimensions and

helps uncover a wide range of features of L2 performance: not only lexical features, but also orthographic, morphological and syntactic aspects.

Despite its potential in revealing neglected aspects of second language writing, our study is not free from certain limitations. One key feature of our research design – the use of a reference corpus to assess the collocational value of L2 bigrams – is double-edged: it is a very rich source of information on the quality of L2 writing, but the value of the results crucially depends on the fit between the reference corpus and the objectives of the study. We opted for the COCA, a large corpus covering a wide range of language uses (speech, fiction, popular magazines, newspapers, academic texts, etc.). Previous studies also used this type of corpus – BNC for Durrant and Schmitt (2009) and Granger and Bestgen (in press) – which has the combined advantage of being very large and representing a wide cross-section of present-day English, two features that make them both good candidates as reference corpora. However, one of the main desiderata for future research will be to assess the respective values of different types of reference corpora. Comparing CollGram outputs based on corpora representing different modes (speech vs. writing) and text types (argumentative vs. narrative) would allow a deeper understanding of learners' writing development, such as a shift from a more personal to a more informational style, as observed by Friginal and Weigle (this issue). This variationist perspective has been recently implemented in some lexical profiling studies (e.g., Lindqvist, Gudmundson, & Bardel, 2013) and ties in closely with Crossley and McNamara's discussion of the norm – written or spoken – used by raters.

Another limitation of the present study is that it is focused exclusively on 2-word sequences, whereas sequences of 3 or 4 words or more have proved to be a particularly good basis for phraseological studies (Biber et al., 2004; Ellis et al., 2010; Ruegg et al., 2011). An additional reason for investigating these longer sequences is that many bigrams (e.g., *in order*, *order to*) are in fact constituents of longer sequences and would therefore deserve to be

investigated at that level (Lyse & Andersen, 2012). However, an extension of CollGram to include longer sequences comes up against a major problem: association measures for *n*-grams of more than two words have been much less investigated and are still on the agenda for future research (Evert, 2009). Biber (2009) provided an extended discussion of this issue. He denounced the common practice of applying to longer sequences the same method to compute MI scores as that used for bigrams. The issue is clearly highly complex, but deserves to be addressed in future versions of CollGram.

To sum up, this study is a first step in assessing the role played by phraseological competence in the development of L2 writing proficiency and text quality assessment by means of an automated technique and quantitative analyses. It confirms and extends a series of recent studies suggesting that phraseology is a key aspect of the study of second language writing (e.g., Henriksen, 2013; Li & Schmitt, 2009; Lindqvist et al., 2013; Vidakovic & Barker, 2010). A great deal of work is required, however, to refine and extend the phraseological analysis and further explore the relationships between phraseological competence and the other linguistic indices used in this special issue to describe L2 writing development.

Acknowledgements

We would like to express our sincere thanks to the journal editors, the editors of the special issue and the anonymous reviewers for the highly valuable feedback they provided on an earlier version of this article. We also gratefully acknowledge the financial support of the Belgian National Fund for Scientific Research (F.R.S.-FNRS).

References

Alsakran, R. A. (2011). *The productive and receptive knowledge of collocations by advanced Arabic-speaking ESL/EFL learners* (Master's thesis). Colorado State University, Fort Collins, Colorado.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*, 275-311.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London, UK: Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). "If you look at...": Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371-405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Bulté & Housen (this issue)

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*, 30-49.

Church, K., & Hanks P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, *16*, 22-29.

Church, K., Gale, W. A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In Uri Zernik (Ed.), *Lexical acquisition: Using on-line resources to build a lexicon*, (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum.

Connor-Linton, J & Polio, C. (this issue)

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics*, (pp. 3168-3171). Oxford, UK: Oxford University Press.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, *16*, 129-147.

Crossley & McNamara (this issue)

De Cock, S. (2000). Repetitive phrase chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory: Papers from the twentieth international conference on English language research on computerized corpora* (pp. 51-68). Amsterdam, The Netherlands: Rodopi.

De Cock, S. (2007). Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight. In M. C. Campoy & M. J. Luzón.(Eds.), *Spoken corpora in applied linguistics* (pp. 217-233). Bern, Switzerland: Peter Lang.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London, UK: Longman.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL: International Review of Applied Linguistics in Language Teaching, 47*, 157-177.

Ebeling, S. O., & Hasselgård, H. (in press). Phraseology in learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of learner corpus research.* Cambridge, UK: Cambridge University Press.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143-188.

Ellis, N. (2003). Construction, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Malden, MA: Blackwell.

Ellis, N., Simpson-Vlach, R., Römer, U., Brook O'Donnell, M., & Wulff, S. (in press). Learner corpora and formulaic language in SLA. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of learner corpus research.* Cambridge, UK: Cambridge University Press.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford, UK: Oxford University Press.

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations* (Doctoral dissertation). University of Stuttgart, Stuttgard, Germany.

Evert, S. (2009). Corpora and collocations. In A. Ludeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1211-1248). Berlin, Germany: Mouton de Gruyter.

Friginal & Weigle (this issue)

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies* (pp. 37-51). Lund Studies in English 88. Lund, Sweden: Lund University Press.

Granger, S., & Bestgen, Y. (in press). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. To appear in *IRAL: International Review of Applied Linguistics in Language Teaching.*

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The international corpus of learner english. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21-33). Houndmills, UK: Palgrave Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development - a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *Vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis* (pp. 29-56). Eurosla Monographs Series 2. Retrieved from http://eurosla.org/.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.

Ishikawa, S. (2009). Phraseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. In K. Yagi & T. Kanzaki (Eds.), *Phraseology, corpus linguistics and lexicography: Papers from phraseology 2009 in Japan* (pp. 87-100). Nishinomiya, Japan: Kwansei Gakuin University Press.

Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. native speakers. *KaLBOTYRa, 61*, 61-72.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, UK: Language Teaching Publications.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*, 85-102.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 Vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109-126). Eurosla Monographs Series 2. Retrieved from http://eurosla.org/.

Lyse, G., & Andersen, G. (2012). Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text. In G. Andersen (Ed.), *Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian*, (pp. 79-110). Amsterdam, The Netherlands: Benjamins.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospects, 16*, 5-19.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford, UK: Oxford University Press.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130-149.

Polio & Shea (this issue)

Ruegg, R., Fritz, E. & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*, 63-80.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics*, *31*, 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*, 510-532.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*, 103-118.

Van Rooy, B., & Schäfer, L. (2003). Automatic POS tagging of a learner corpus: the influence of learner errors on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.). *Proceedings of the corpus linguistics 2003 conference* (pp. 835-844). Technical Papers 16. Lancaster, UK: Lancaster University Centre for Computer Corpus Research on Language.

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life Writing Examinations. *Cambridge ESOL: Research Notes, 41*, 7-14.

Weltig, M. S. (2004). Effects of Language Errors and Importance Attributed to Language on Language and Rhetorical-Level Essay Scoring. *Spaan fellow working papers in*

second or foreign language assessment. Volume 2 (pp. 52-79). Ann Arbor, MI: English Language Institute, University of Michigan.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*, 647-668.

Table 1

Details of the MSU Learner Corpus

	Value
Number of texts	171
Total number of words	57358
Number of words per text	
Mean	335.4
SD	97.7
Minimum	141
Maximum	655
Number of bigrams per text	
Mean	292.1
SD	87.7
Minimum	124
Maximum	585

Table 2

Bigrams and MI Scores for two Excerpts from Learner Texts

Essay 122 Text quality score 70			Essay 296 Text quality score 53.5		
Token	Bigram	MI	Token	Bigram	MI
He			In		
used	he_used	2.06	the	in_the	2.32
to	used_to	3.84	morning	the_morning	1.72
investigate	to_investigate	4.62	everything	morning_everything	-0.51
illegally	investigate_illegally	.	are	everything_are	-3.86
parked	illegally_parked	9.72	peaceful	are_peaceful	0.51
cars	parked_cars	9.08	Nothing	peaceful_nothing	.
.			is	nothing_is	1.17
Now			show	is_show	-3.27
,			to	show_to	-1.56
he			you	to_you	-0.45
came	he_came	3.11	the	you_the	-3.76
here	came_here	4.13	difference	the_difference	2.36
with	here_with	1.78	in	difference_in	3.52
me	with_me	2.13	the	in_the	2.32
in	me_in	0.80	night	the_night	1.59
order	in_order	4.51	.		
to	order_to	4.22	In		
improve	to_improve	4.31	the	in_the	2.32
his	improve_his	2.10	nights	the_nights	-0.61
ability	his_ability	3.28	everything	nights_everything	-1.19
to	ability_to	4.75	are	everything_are	-3.88
speak	to_speak	3.92	changed	are_changed	-0.20

English fluently .	speak_english english_fluently	8.36 9.73	.

Table 3

Top-scoring (Left) and Lowest-scoring (Right) Bigrams (MI and t-score)

Top-scoring bigrams	MI	Top-scoring bigrams	t	Lowest-scoring bigrams	MI	Lowest-scoring bigrams	t
ping_pong	17.7	of_the	1241	his_my	-11.0	the_they	-9415
ha_ha	15.7	in_the	1142	a_out	-10.8	i_the	-7656
rocket_launchers	14.1	on_the	804	the_across	-10.3	a_out	-6821
vacuum_cleaner	13.8	do_not	774	the_they	-10.1	on_of	-4858
alcoholic_beverage	13.8	to_be	740	i_their	-10.1	i_a	-3986
ozone_layer	12.8	that_is	648	been_are	-10.0	i_their	-3757
fire_extinguisher	12.6	at_the	644	i_fact	-9.9	to_had	-3675
korean_peninsula	12.6	it_was	641	our_from	-9.7	a_not	-3602
toxic_substances	12.5	did_not	612	more_many	-9.7	to_in	-3538
grand_rapids	12.3	going_to	599	as_same	-9.6	of_they	-3455
ice_cream	12.2	there_is	580	we_has	-9.4	they_is	-3137
vending_machine	12.2	i_think	554	their_another	-9.4	the_across	-3015
hello_kitty	12.1	from_the	553	their_are	-9.3	his_my	-2930
swimming_pool	12.1	for_the	540	he_special	-9.3	my_and	-2929
microwave_oven	11.9	you_know	524	a_ability	-9.2	the_some	-2784
lung_cancer	11.9	it_is	515	my_another	-9.2	the_see	-2489
soviet_union	11.7	as_a	508	include_of	-9.2	to_can	-2282
personality_traits	11.7	he_was	508	to_entered	-9.1	the_all	-2148
amusement_parks	11.6	in_a	495	they_problems	-9.1	in_not	-2147
two-lane_roads	11.5	and_i	491	he_teachers	-9.0	the_so	-2144
monetary_fund	11.3	one_of	486	every_my	-8.9	to_out	-2083
acid_rain	11.3	he_is	485	a_each	-8.9	been_are	-2045

human_beings	11.3	i_do	475	to_against	-8.9	we_has	-1957
bulletin_board	11.2	with_the	473	in_keep	-8.8	their_are	-1906
credit_card	11.2	we_have	471	filled_of	-8.8	i_and	-1841
convenience_store	11.2	have_been	466	not_difference	-8.8	to_against	-1708
jung_hee	11.1	to_the	466	which_be	-8.7	a_each	-1644
marching_bands	11.1	with_a	462	they_is	-8.7	to_said	-1619
ice_rink	11.0	can_not	460	to_went	-8.7	of_out	-1618
traffic_jam	11.0	by_the	460	my_and	-8.6	this_the	-1562
household_appliances	10.9	this_is	460	a_experience	-8.6	i_has	-1528
roller_skate	10.8	out_of	458	a_foods	-8.6	i_one	-1487
diesel_engine	10.7	is_a	457	to_had	-8.4	our_from	-1435
ice_crystals	10.7	want_to	457	other_my	-8.4	to_went	-1410
academic_achievement	10.6	if_you	454	if_make	-8.3	more_not	-1383
committed_suicide	10.5	more_than	444	of_they	-8.3	you_the	-1362
middle_east	10.4	does_not	439	all_person	-8.2	is_of	-1341
dorm_rooms	10.4	i_was	439	to_came	-8.2	i_fact	-1320
ethernet_cable	10.4	the_same	435	on_of	-8.2	their_i	-1278
dining_room	10.6	a_lot	433	they_has	-8.1	not_has	-1273
sun_shines	10.3	has_been	428	he_son	-8.1	which_be	-1265
vice_president	10.2	the_first	425	their_i	-8.1	they_has	-1261
touristic_attraction	10.2	is_not	423	a_biggest	-8.1	to_came	-1242
international_monetary	10.1	will_be	419	during_i	-8.0	of_or	-1153
hazardous_materials	10.0	into_the	414	she_teachers	-8.0	as_same	-1131
rental_fee	10.0	would_be	413	the_see	-8.0	with_in	-1121
baseball_bats	10.0	to_get	411	huge_is	-8.0	in_keep	-1110
slot_machine	10.0	to_do	408	i_the	-8.0	a_many	-1109
off-campus_apartment	10.0	for_a	407	he_future	-8.0	the_seen	-1072

rainy_season	9.9	he_said	393	an_higher	-7.9	to_there	-1033

Table 4

Examples of Erroneous Bigrams in the Absent Category

Line	Bigram	Context
1	be droved	buses can be droved into campus
2	he beguns	he beguns to use the machine
3	every facilities	it has every facilities
4	public transportations	the government should promote citizens to use public transportations
5	be housewife	she should just be housewife
6	has tendency	a good teacher has tendency for having many students
7	likes talk	he likes talk with everybody
8	arrived campus	As I arrived campus in Fall 1999
9	resemble with	He resemble with me
10	graduate their	After graduate their school, they can make money
11	always responses	he always responses every email immediately
12	they establishment	If they establishment a kind of equipment that make the smoke harmless
13	had pregnated	In fact, she had pregnated 7 times
14	scarly and	It was very scarly and interested to me

Table 5:

Change in Collgram Scores Over the Semester

	Start		End	
	M	SD	M	SD
MI tokens	2.16	0.24	2.10	0.25
MI types	2.02	0.25	1.97	0.26
t-score tokens	105.48	20.25	97.99	19.09
t-score types	84.76	17.33	78.34	14.84
Prop. Absent Tokens	0.03	0.02	0.03	0.02
Prop. Absent Types	0.04	0.02	0.03	0.02

Table 6

Correlations Between the Three Indices (Tokens and Types) and Text Quality Ratings

	Scale		
	Combined	Language	Vocabulary
MI tokens	0.28***	0.32***	0.22**
MI types	0.35***	0.43***	0.31***
t-score tokens	0.11	0.14	0.08
t-score types	0.03	0.10	0.02
Prop. Absent Tokens	-0.27***	-0.36***	-0.15*
Prop. Absent Types	-0.28***	-0.37***	-0.16*

*$p < .05$. **$p < .01$. ***$p < .001$.